Assignment 11

Christophe Hunt

April 23, 2017

Contents

Using R's lm function, perform regression analysis and measure the significance of the independent variables for the following two data sets.

In the first case, you are evaluating the statement that we hear that Maximum Heart Rate of a person is related to their age by the following equation:

MaxHR = 220 - Age

You have been given the following sample:

Age 18 23 25 35 65 54 34 56 72 19 23 42 18 39 37
MaxHR 202 186 187 180 156 169 174 172 153 199 193 174 198 183 178

Perform a linear regression analysis fitting the Max Heart Rate to Age using the lm function in R. What is the resulting equation?

```
library(stargazer)
Age <- c(18, 23, 25, 35, 65, 54, 34, 56, 72, 19, 23, 42, 18, 39, 37)
MaxHR <- c(202, 186, 187, 180, 156, 169, 174, 172, 153, 199, 193, 174, 198, 183, 178)
fit <- lm(MaxHR~Age)
stargazer(fit, header = FALSE)
```

Table 1:

	Dependent variable:
	MaxHR
Age	-0.798^{***}
	(0.070)
Constant	210.048^{***}
	(2.867)
Observations	15
R^2	0.909
Adjusted R^2	0.902
Residual Std. Error	4.578 (df = 13)
F Statistic	130.009^{***} (df = 1; 13)
Note:	*p<0.1; **p<0.05; ***p<0.01

Is the effect of Age on Max HR significant?

> Yes, the effect of Age on MaxH is significant. We see the p<0.01 for Age which indicates that it is indeed signficiant. We would say anything with a p<0.05 is typically significant, depending on the situation.

What is the significance level?

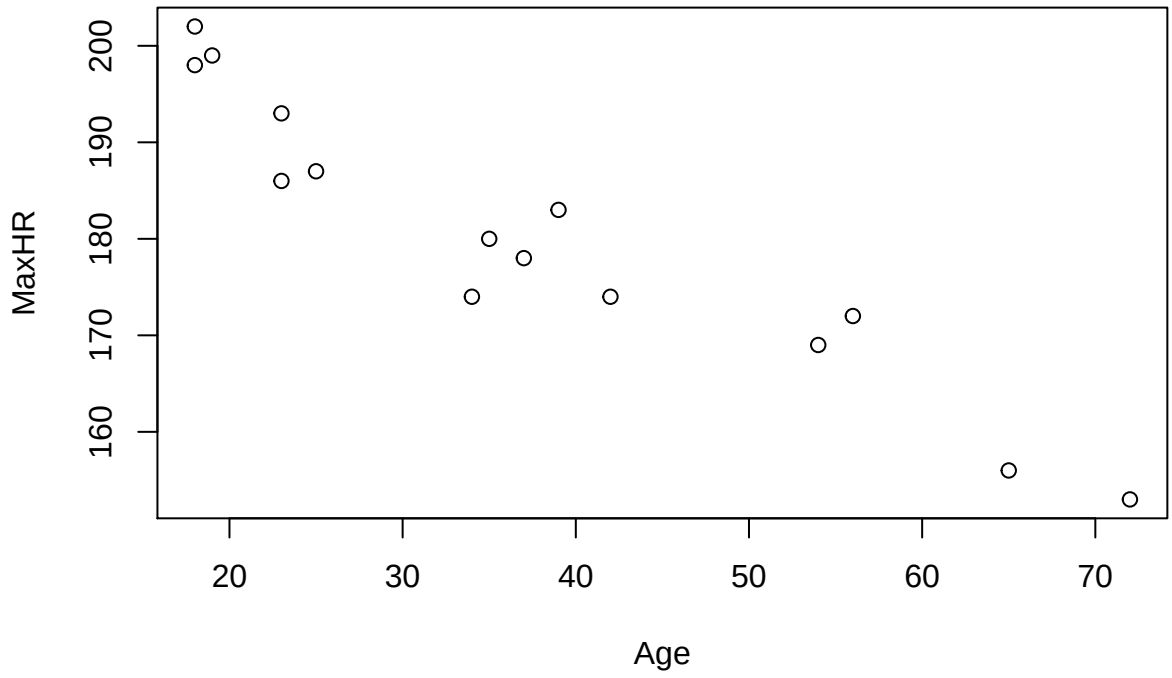
The signficance level is p<0.01, the exact value is as follows.

```
options(scipen = 999)
x <- anova(fit)
x$`Pr(>F)`[1]
```

```
## [1] 0.00000003847987
```

Please also plot the fitted relationship between Max HR and Age.

```
plot(MaxHR~Age)
```



Using the Auto data set from Assignment 5 (also attached here) perform a Linear Regression analysis using mpg as the dependent variable and the other 4 (displacement, horse-power, weight, acceleration) as independent variables.

What is the final linear regression fit equation?

```
library(tidyverse)
df <- as_tibble(read.table(paste0("https://raw.githubusercontent.com",
                                  "/ChristopheHunt/MSDA---Coursework",
                                  "/master/Data%20605/Assignment%2011/",
                                  "auto-mpg.data")))
colnames(df) <- c("displacement", "horsepower", "weight", "acceleration", "mpg")
```

```
fit <- lm(data = df, formula = (mpg ~ displacement +  horsepower +
                                  weight + acceleration))
stargazer(fit, header = FALSE)
```

Table 2:

	Dependent variable:
	mpg ~displacement + horsepower + weight + acceleration
displacement	−0.006
	(0.007)
horsepower	−0.044***
	(0.017)
weight	−0.005***
	(0.001)
acceleration	−0.023
	(0.126)
Constant	45.251***
	(2.456)
Observations	392
R^2	0.707
Adjusted R^2	0.704
Residual Std. Error	4.247 (df = 387)
F Statistic	233.434*** (df = 4; 387)
Note:	*p<0.1; **p<0.05; ***p<0.01

Which of the 4 independent variables have a significant impact on mpg?

> Horsepower and weight have a significant impact on mpg. It is unlikely that their values have an impact on mpg merely by chance.

What are their corresponding significance levels?

```
summary(fit)[4]
```

```
## $coefficients
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  45.251139699 2.4560446927 18.4243959 7.072099e-55
## displacement -0.006000871 0.0067093055 -0.8944102 3.716584e-01
## horsepower   -0.043607731 0.0165734633 -2.6311779 8.848982e-03
## weight       -0.005280508 0.0008108541 -6.5122789 2.302545e-10
## acceleration -0.023147999 0.1256011622 -0.1842977 8.538765e-01
```

> The variables corresponding significance levels are horsepower = 0.009 & weight = 0.0000000002

What are the standard errors on each of the coefficients?

```
summary(fit)[4]$coefficients
```

```
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  45.251139699 2.4560446927 18.4243959 7.072099e-55
## displacement -0.006000871 0.0067093055 -0.8944102 3.716584e-01
## horsepower   -0.043607731 0.0165734633 -2.6311779 8.848982e-03
## weight       -0.005280508 0.0008108541 -6.5122789 2.302545e-10
## acceleration -0.023147999 0.1256011622 -0.1842977 8.538765e-01
```

> The variables corresponding standard errors for our coefficients are horsepower = 0.0166 & weight = 0.00081

Please perform this experiment in two ways.

First take any random 40 data points from the entire auto data sample and perform the linear regression fit and measure the 95% confidence intervals.

```
set.seed(1234)
df_sample <- sample_n(df, 40)
fit.samp <- lm(data = df_sample, formula = (mpg ~ displacement +
                                              horsepower + weight + acceleration))
stargazer(fit.samp, header = FALSE)
```

Table 3:

	Dependent variable:
	mpg ~displacement + horsepower + weight + acceleration
displacement	0.010
	(0.030)
horsepower	−0.177**
	(0.078)
weight	−0.002
	(0.004)
acceleration	−0.413
	(0.498)
Constant	52.995***
	(8.262)
Observations	40
R^2	0.710
Adjusted R^2	0.677
Residual Std. Error	4.595 (df = 35)
F Statistic	21.396*** (df = 4; 35)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
names <- colnames(df)[1:4]
for (i in names){
kable(print(confint(fit.samp, i, level = 0.95)))
}
```

```
##                   2.5 %     97.5 %
## displacement -0.0509389 0.07008803
##                  2.5 %      97.5 %
## horsepower -0.3361474 -0.01804589
##                2.5 %      97.5 %
## weight -0.01142215 0.006696022
##                   2.5 %   97.5 %
## acceleration -1.423285 0.596921
```

Then, take the entire data set (all 392 points) and perform linear regression and measure the 95% confidence intervals.

```
fit <- lm(data = df, formula = (mpg ~ displacement +
                                  horsepower + weight + acceleration))
stargazer(fit, header = FALSE)
```

Table 4:

	Dependent variable:
	mpg ~displacement + horsepower + weight + acceleration
displacement	−0.006
	(0.007)
horsepower	−0.044***
	(0.017)
weight	−0.005***
	(0.001)
acceleration	−0.023
	(0.126)
Constant	45.251***
	(2.456)
Observations	392
R^2	0.707
Adjusted R^2	0.704
Residual Std. Error	4.247 (df = 387)
F Statistic	233.434*** (df = 4; 387)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
names <- colnames(df)[1:4]
for (i in names){
kable(print(confint(fit, i, level = 0.95)))
}
```

```
##                    2.5 %     97.5 %
## displacement -0.01919212 0.00719038
##                   2.5 %      97.5 %
## horsepower -0.07619303 -0.01102243
##                2.5 %       97.5 %
## weight -0.006874738 -0.003686277
##                   2.5 %    97.5 %
## acceleration -0.270094 0.2237981
```

Let's compare the results of the

```
names <- colnames(df)[1:4]
for (i in names){
kable(print(confint(fit, i, level = 0.95)))
kable(print(confint(fit.samp, i, level = 0.95)))
}
```

```
##                    2.5 %     97.5 %
## displacement -0.01919212 0.00719038
```

```
##                   2.5 %     97.5 %
## displacement -0.0509389 0.07008803
##                   2.5 %      97.5 %
## horsepower -0.07619303 -0.01102243
##                  2.5 %      97.5 %
## horsepower -0.3361474 -0.01804589
##                2.5 %       97.5 %
## weight -0.006874738 -0.003686277
##               2.5 %      97.5 %
## weight -0.01142215 0.006696022
##                   2.5 %    97.5 %
## acceleration -0.270094 0.2237981
##                   2.5 %   97.5 %
## acceleration -1.423285 0.596921
```

Please report the resulting fit equation, their significance values and confidence intervals for each of the two runs.

Please submit an R-markdown file documenting your experiments. Your submission should include the final linear fits, and their corresponding significance levels.

In addition, you should clearly state what you concluded from looking at the fit and their significance levels.